

Human-chimpanzee alignment: Ortholog Exponentials and Paralog Power Laws

Kun Gao and Jonathan Miller

Physics and Biology Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan

Abstract

Genomic subsequences conserved between closely related species such as human and chimpanzee exhibit an exponential length distribution, in contrast to the algebraic length distribution observed for sequences shared between distantly related genomes. We find that the former exponential can be further decomposed into an exponential component primarily composed of orthologous sequences, and a truncated algebraic component primarily composed of paralogous sequences.

Keywords: evolution, genomic alignment, length distribution, exponential and power-law, orthology and paralogy

1. Introduction

Sequence conservation is defined by similar or identical nucleotide sequences within or among genomes at frequencies beyond those expected on neutral evolution. Within most neutral models of evolution, the probability that a sequence appears in two unrelated genomes decays exponentially with its length, so that sufficiently long sequences common to more than one genome are expected to derive from a common ancestor [15]. Sequence duplication represents a primary mechanism through which new genetic material can arise [23, 28]. When identical sequences are observed within a single genome at levels exceeding those expected on an independent site model of evolution, sequence duplication is one candidate for their origin. Similarity among sequences beyond that expected within an independent site model, whether multiple occurrences within a single genome or simultaneous occurrence in multiple genomes, is known as “sequence homology” and may

indicate common ancestry [4].

Because sequence conservation and sequence duplication are often inferred from sequence length and identity, we believe that a systematic understanding of the latter two features may elucidate rules underlying sequence evolution and lead to more faithful models of neutral evolution.

The set of sequences shared within a genome or between two genomes may be summarised in its “length distribution:” a histogram with length L on the x -axis and number $\#(L)$ of shared sequences of length L on the y -axis. These length distributions can exhibit distinctive characteristics that we aim to account for within some model of sequence evolution. Henceforth, we abbreviate “length distribution” to ‘distribution,’ as all distributions referred to in this manuscript are histograms of the form indicated above.

Strong conservation among distantly related genomes: algebraic distribution with exponent ≈ -4

Distributions of sequences strongly conserved between a variety of distantly related genome pairs exhibit a heavy, approximately algebraic (power-law) tail [30]. This power-law distribution is common but not universal; occurs not merely pairwise but also among multiple genomes; is robust over different measures of similarity; and has an exponent reported to be typically in the neighborhood of -4 . Thus “ultra-conserved” sequences exhibit this power law, but the same exponent also governs pairwise conserved sequences that are not necessarily shared by a third genome.

Sequence identity among closely related genomes: exponential distribution

Sequences conserved between closely related species such as human and certain primates display an exponential distribution, rather than a power-law [26]. “Closely related” is defined empirically as sufficiently recent branching from a common ancestor.

“Ultra-duplication” within single genomes: algebraic distribution with exponent -3

Study of exact duplications of all lengths in different genomes through whole-genome/whole-chromosome self-alignment revealed that duplicates with 100% identity often – but not always – follow an approximately algebraic distribution with exponent in the neighbourhood of -3 [9]. Since it was originally observed (although with different exponent) in ultra-conserved sequences, in the context of duplicated sequences this algebraic feature was

referred to as “ultra-duplication,” the prefix “ultra” alluding solely to the long tail of the corresponding distribution, irrespective of its origin.

Massip and Arndt recently observed that together segmental duplication and point mutation can yield an algebraic distribution with exponent -3 [24] (see also [17]). Customarily, segmental duplication is thought of as a neutral process, although selection may act subsequently.

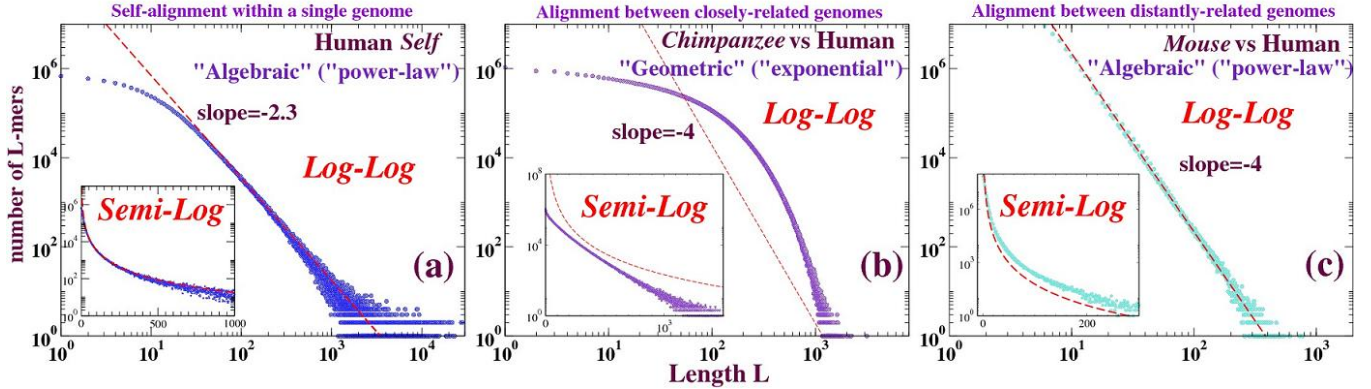


Figure 1: Distributions of identically conserved or duplicated sequences in (a) human chromosome 1 self-alignment; (b) human chromosome 1 – chimpanzee chromosome 1 alignment and (c) human chromosome 1 – mouse chromosome 1 alignment. Distributions shown are generated by repeat-masked whole-chromosome LASTZ [13] net alignments obtained directly from UCSC Genome Bioinformatics. Log-log plots enclose semi-log insets.

Figure 1 recapitulates the three cases mentioned above. With increasing evolutionary distance between species, distributions of identical sequences obtained from LASTZ net alignment cross over from algebraic (with exponent -3) to exponential, and then again to algebraic (with exponent in the neighborhood of -4). These crossovers are further elucidated below. All alignments described in this manuscript were performed with LASTZ (see **Materials and Methods**); henceforth – with the exception of the “**Materials and Methods**” section – we refer to “alignment” and for the most part we omit the qualifier “LASTZ,” which is tacitly implied unless otherwise indicated explicitly.

In the following, we apply whole-genome/whole-chromosome alignment between human and chimpanzee to investigate the origin of the exponential distribution and disentangle it from the algebraic distribution. For closely

related species, quantitative relationships emerge between orthologous sequences and the exponential distribution, and between paralogous sequences and the algebraic distribution.

2. Materials and Methods

2.1. Pairwise alignment of genome sequences

Software

We compare genomic sequences with the LASTZ pairwise alignment tool [13]. LASTZ alignment comprises several stages of which we rely mainly on two: *raw* alignment and *net* alignment. Raw alignment is the immediate product of LASTZ and may include multiple and positionally overlapping matches for each aligned sequence. A subsequent net alignment removes positional overlaps among matched sequences, chains them, and discards all but the highest-scoring chains, yielding a single match for each position in the genome. One function of net alignment is to extract homologous elements from the raw alignment [14].

LASTZ is obtained from http://www.bx.psu.edu/miller_lab/; we use LASTZ default options for raw alignment. The UCSC Genome Browser (<http://hgdownload.cse.ucsc.edu/admin/jksrc.zip>) provides additional tools (*axtChain*, *chainNet* and *netToAxt*) for producing the net alignment. Standard procedures that we follow for LASTZ alignment (both raw and net) are described at: http://genomewiki.cse.ucsc.edu/index.php/Whole_genome_alignment_howto.

Genome sequences

Soft repeat-masked (<http://repeatmasker.org>) genome sequences are obtained as fasta files from the Ensembl FTP Server (e.g. *hg19* as version 74; ftp://ftp.ensembl.org/pub/release-74/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.74.dna.chromosome.1.fa.gz). We use for the most part the *hg19* and *panTro4* assemblies for human and chimpanzee respectively. For most of our calculations, we study the human chromosome 1 – chimpanzee chromosome 1 LASTZ raw alignment.

Other primate genomes are obtained from Ensembl, and human chromosome 1 is aligned to its respective “orthologous” counterpart from each primate – the primate chromosome that shares with human chromosome 1 the most orthologous genes as identified in Ensembl Biomart (<http://www.ensembl.org/biomart/martview>), yielding gorilla chromosome 1, orangutan

chromosome 1 (reverse strand), macaca chromosome 1, and marmoset chromosome 7 as orthologous to human chromosome 1.

Mouse (*mm9*) chromosome 1 is downloaded from Ensembl and aligned to human chromosome 1. Mouse chromosome 1 carries a plurality (close to 1/4) of orthologous elements shared between human chromosome 1 and the mouse whole genome. The Venter genome [21] is obtained from UCSC Genome Bioinformatics (<http://hgdownload.cse.ucsc.edu/goldenPath/venter1/bigZips/venter1.2bit>) and aligned to *hg19*.

Repeat-masking

Repetitive sequence elements may constitute close to half the genome; unless these repeats are explicitly identified, most if not all large-scale alignment methods may fail to complete on eukaryotic genomes. A common practice is to identify these elements prior to alignment with a software tool such as Repeatmasker (<http://repeatmasker.org>) that demarcates repetitive sequence in lower-case letters (“soft [repeat-]masking”) in contrast to the upper-case letters that designate non-repetitive sequence.

Unless otherwise indicated, all LASTZ alignments represented here are performed on soft repeat-masked genome sequences.

When aligning sequences, LASTZ excludes soft-masked bases from its “seed” stage but reintroduces them in later stages, when alignments of unmasked sequence can be extended into soft-masked regions (see webpage: http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html#overview), so that LASTZ can in principle align certain masked sequences. For example, just over 50% of each of human (*hg19*; Ensembl GRCh37.74) chromosome 1 and chimpanzee (Ensembl CHIMP2.1.4.74) chromosome 1 is repeat-masked. Nevertheless, the LASTZ raw alignment between these two soft repeat-masked chromosomes covers 94% of the human chromosome, and 97% of the chimpanzee chromosome; 92% of the masked bases in human chromosome 1 and 96% of the masked bases in chimpanzee chromosome 1 are aligned by LASTZ.

2.2. Parsing the alignment

Distribution of contiguous matched runs (CMRs)

Following [9], for a given pairwise alignment we study continuous (uninterrupted) matching runs of bases (CMRs), wherein a contiguous series of matching nucleotides is terminated at mismatches or indels. Unless otherwise

indicated, all CMRs discussed here represent exact matches that we refer to interchangeably with these two terms.

A histogram $\#(L)$ (or (length) distribution) describes the number of CMRs of a given length L . Pairwise alignment of genomes yields conserved or duplicated sequences within or between genomes; we expect that distributions of these conserved or duplicated sequences reflect certain features of genome evolution.

Forward alignment and reverse alignment

DNA is composed of complementary strands so that for two DNA sequences pairwise alignment can be implemented in either of two relative orientations, “forward” or “reverse.” Matches to the reverse orientation are thought to arise by inversion or inverted duplication/transposition [1]. We perform both forward and reverse alignments; however, we subsequently combine their products before further calculation except where it is informative to keep them separate.

Dot plot

A two-dimensional similarity matrix between two sequences is displayed as a dot plot [10], in which one sequence of an aligned pair lies along the horizontal and the other along the vertical axis. Dot plots are commonly used to visualise sequence similarity and to display homologous matches between genomes. “Syntenic dot plots” exhibit synteny (see webpage: http://genomeevolution.org/wiki/index.php/Syntenic_dotplots). In this paper, we apply them to display orthologous sequences between human and chimpanzee genomes.

In some of our dot plots, prominent horizontal or vertical white bands appear that correspond to sequence that has not yet been reliably determined and is therefore represented by “N” in the assemblies from Ensembl; such bases are excluded from alignments [13].

3. Results

3.1. Alignments of orthologous regions of genomes yield exponential distributions

An alignment of a numbered human chromosome to the chromosomes of chimpanzee yields the exponential distribution of CMRs shown in figure 1 for the correspondingly numbered chimpanzee chromosome, but not for

other chimpanzee chromosomes. In the latest releases of the chimpanzee genome assembly, chimpanzee chromosomes have been renumbered to reflect common ancestry with the corresponding human chromosomes [25], so that chromosomes sharing the same number can be thought of as “orthologous chromosomes.” In figure 2, human chromosome 1 is separately aligned against each chimpanzee chromosome. With the exception of the alignment of human chromosome 1 with chimpanzee chromosome 1, all the alignments (raw and net) between human chromosome 1 and chimpanzee chromosomes yield approximately algebraic distributions. The dot plots corresponding to these alignments can be found in figure *S1*.

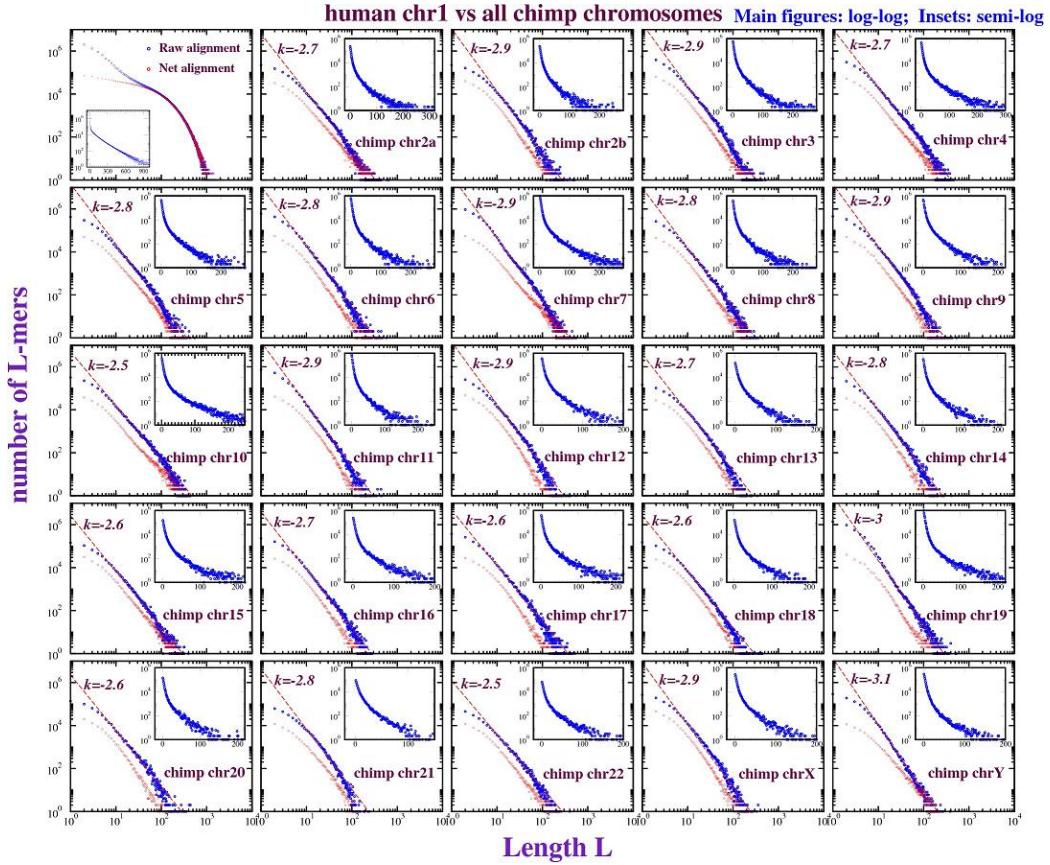


Figure 2: Distributions of exact matches (CMRs) in raw (blue) and net (red) alignments of human chromosome 1 against all chimpanzee chromosomes. Main figures show log-log plots; insets semi-log plots for the same data. For purposes of comparison, lines with slope k on the log-log scale as indicated have been drawn into each figure.

Although human chromosome 1 – chimpanzee chromosome 1 alignment exhibits an exponential distribution overall, figure 3 suggests that this exponential is composed solely or primarily of CMRs between orthologous regions of these two chromosomes. Figure 3 illustrates alignments of orthologous versus heterologous sequences in human chromosome 1 and chimpanzee chromosome 1. *H_frag1*, *H_frag2*, *C_frag1* and *C_frag2* are fragments taken respectively from the first and last thirds of these two chromosomes. For these four fragments, figure 4 shows an orthology map and figure *S2* a syntenic dot plot. As can be seen in figure 3, alignments between homologous (heterologous) fragments exhibit exponential (algebraic) distributions.

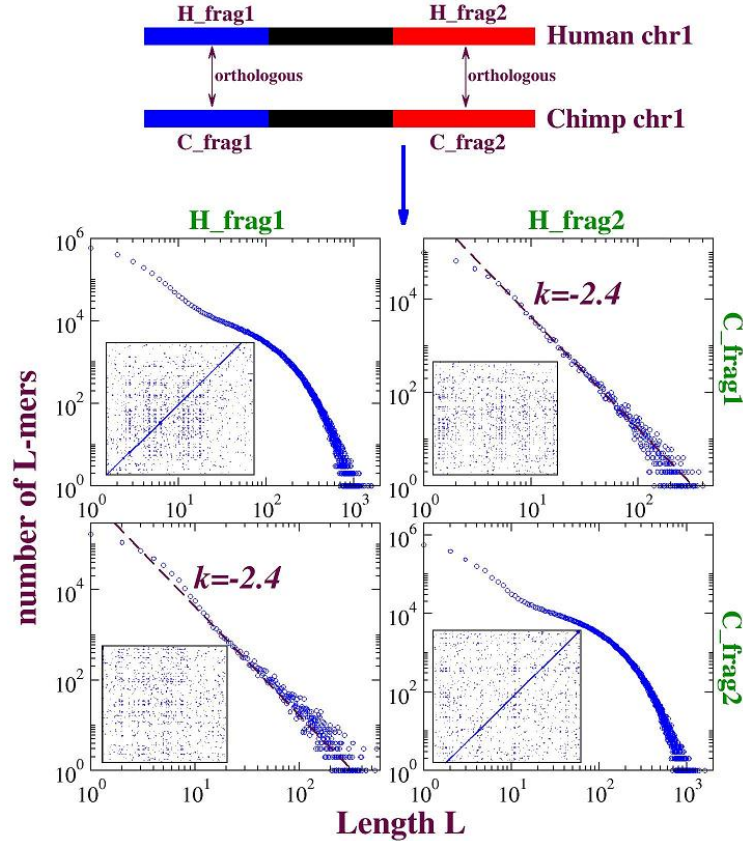


Figure 3: Distributions of exact matches in raw alignments between different fragments of human chromosome 1 and chimpanzee chromosome 1. We extract the first and last thirds

of each these two chromosomes, and align all four resulting fragment pairs. These figures indicate that even within a single chromosome, the exponential distribution correlates with orthology: only alignments between orthologous fragments show exponential distributions.

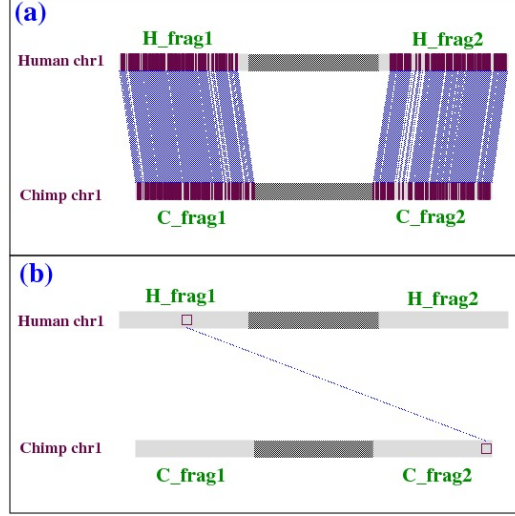


Figure 4: Orthology map among different fragments of human chromosome 1 and chimpanzee chromosome 1. Horizontal dark grey bars (largely obscured in (a) by maroon vertical bars) show human chromosome 1 and chimpanzee chromosome 1; light grey bars indicate fragments H_frag1 , H_frag2 , C_frag1 and C_frag2 defined in figure 3. Maroon vertical bars indicate the locations genes orthologous between human chromosome 1 and chimpanzee chromosome 1; and blue dotted lines connect these orthologous gene pairs.

We describe further evidence below that the exponential distribution observed in the human-chimpanzee alignment correlates with sequence orthology.

3.2. Separating the exponential from the algebraic

Although alignment of orthologous human and chimpanzee chromosomes yields an exponential distribution overall, distributions of aligned subfragments of these genomes are not necessarily exponential. In figures 2 and 3 it is seen that whole-chromosome alignments between human and chimpanzee contain both exponential and power-law components.

In this subsection, we illustrate how the human chromosome 1 – chimpanzee chromosome 1 alignment naturally decomposes into algebraic and exponential subsets. Based on the observations in figure 2 and figure 3, we hypothesise that

to a first approximation, the exponential and (approximately) power-law components correspond to orthologous sequence and paralogous sequence, respectively.

To perform this decomposition we develop several methods, each of which is related to this hypothesis in a slightly different way. With the exception of a “local” method based on “nested” and “non-nested” matches that is parameter-free, they involve parameter choices and sometimes further manipulations whose justification is not always readily apparent. Nevertheless, it turns out that these methods yield very similar outcomes.

It may be worth remarking that a length distribution alone contains no information about location in a genome, so that it is impossible to partition an alignment into exponential and algebraic components solely on the basis of aligned fragment or CMR lengths; nevertheless, the content of the previous subsection suggests that a partition can be extracted from the dot plot.

3.2.1. “Geometrical” method: Separating on-diagonal elements from off-diagonal elements

As evident from figure 2, figure 3 and figure *S1*, alignments between human and chimpanzee genomes with an exponential distribution exhibit dense accumulations of sequences within the dot plot. For closely related species like human and chimpanzee, it is well known that one such high density zone ordinarily forms a band near the diagonal of the dot plot that we refer to as the “diagonal band.” We will see that the diagonal band is a major contributor to the exponential distribution.

Figure 5 shows the distributions of exact matches from the diagonal band and from its complement within the dot plot of the human chromosome 1 – chimpanzee chromosome 1 raw alignment. We crudely take into account the length difference between human chromosome 1 and chimpanzee chromosome 1, on the order of $\delta_C1 = 10^7$ bases, by defining a region around the diagonal of width δ_C1 , so that sequences offset by as many as δ_C1 bases are to be thought of – in this approximation – as comprising the diagonal band. The exponential distributions of CMRs in this diagonal band and the algebraic distributions of the off-diagonal CMRs are evident in the right-most panels of figure 5.

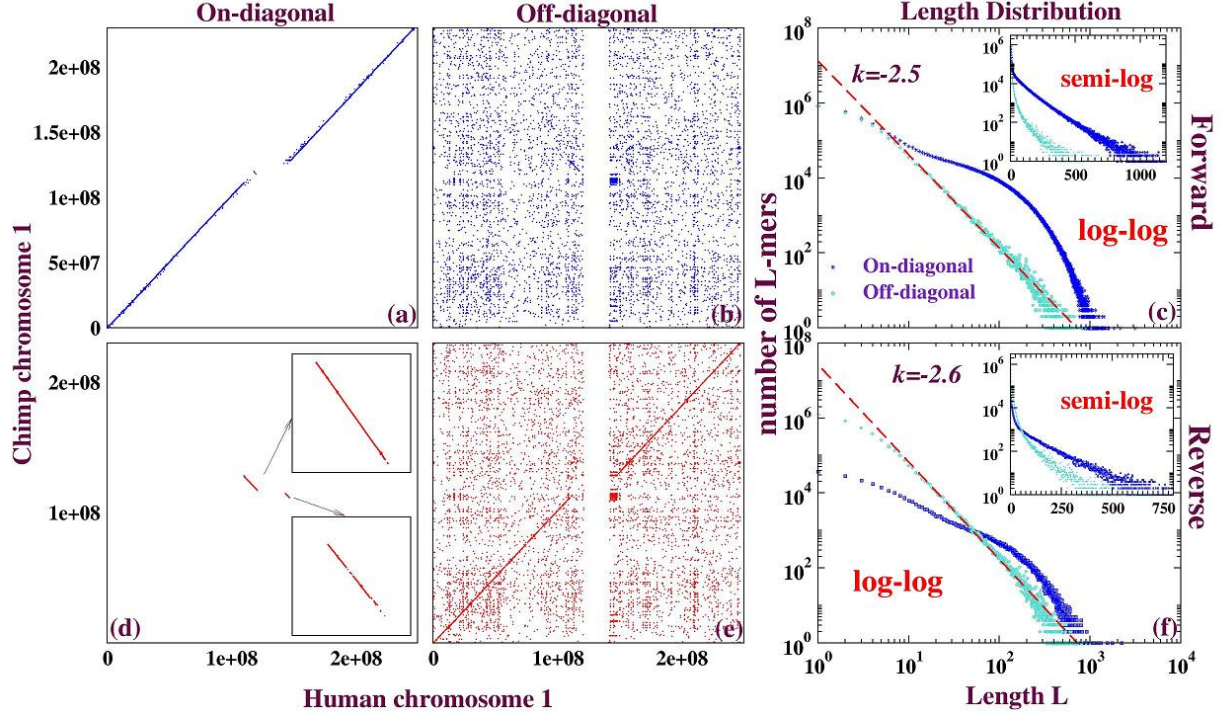


Figure 5: Dot plots and distributions of exact matches on diagonal and off diagonal in human chromosome 1 – chimpanzee chromosome 1 raw alignment. The diagonal band width is chosen as $\delta_C 1 = 10^7$ bases (see text). For the reverse alignment, we excise not the whole diagonal band, but rather only two fragments (see the insets in the lower-left panel) that in the dot plot correspond to large inversions. The exponential distribution of CMRs in this diagonal band and the algebraic distribution of off-diagonal CMRs is evident in the right-most panels.

In contrast to the other methods described here, in this subsection we treat the forward and reverse alignment separately. As we have discussed above, the orthologous sequences between human chromosome 1 and chimpanzee chromosome 1 concentrate primarily in the forward strands; we extract the entire diagonal band from the forward alignment. In the reverse alignment, two large and distinct inversions appear on the dot plot (see insets in lower-left panel of Figure 5); by extracting these inversions we find empirically that we can neatly separate exponential from power-law in the reverse alignment. This can be understood if these large inversions are recent events, so that in contrast to the rest of these two chromosomes, the orthologous orientation is reversed.

3.2.2. “Genetic clock” method: Separating high-similarity alignment blocks from low-similarity alignment blocks

The raw alignment is composed of a set of alignment blocks, each representing a local alignment whose score is higher than a pre-established threshold. One way to characterise similarity in an alignment block is to compute the number of mismatches it contains, yielding a Hamming distance. The ratio of Hamming distance to sequence length then represents a (time-integrated) rate of variation per base. In the absence of selection, and under customary idealisations, this ratio reflects the time elapsed subsequent to divergence, constituting a crude “genetic clock” [34]. According to the definition of ortholog and paralog (see section 4.1), paralogs diverge before orthologs and should exhibit greater ratios of Hamming distance to sequence length than orthologs. Figure 6 left panel shows Hamming distance versus alignment block length for all alignment blocks in the human chromosome 1 – chimpanzee chromosome 1 raw alignment; each dot corresponds to an alignment block. Evidently, these alignment blocks comprise two major branches; alignment blocks in the lower-right branch exhibit lower rates of variation (greater similarity) than those in the upper-left branch.

Figure 6 left panel shows a natural partition of alignment blocks; a line with slope in the neighborhood of 0.08 from the origin is sufficient to elucidate a partition into an upper-left branch and a lower-right branch (leftmost panel). Middle panels in figure 6 show respective dot plots for the alignment blocks in the upper-left branch and lower-right branch, and the rightmost panel the (approximately) algebraic distribution of the upper-left branch and the exponential distribution of the lower-right branch.

In this “genetic clock” method, we treat the forward and reverse alignments identically, thus in figure 6, we display the combined product of forward and reverse alignment; they are exhibited separately in figure S3.

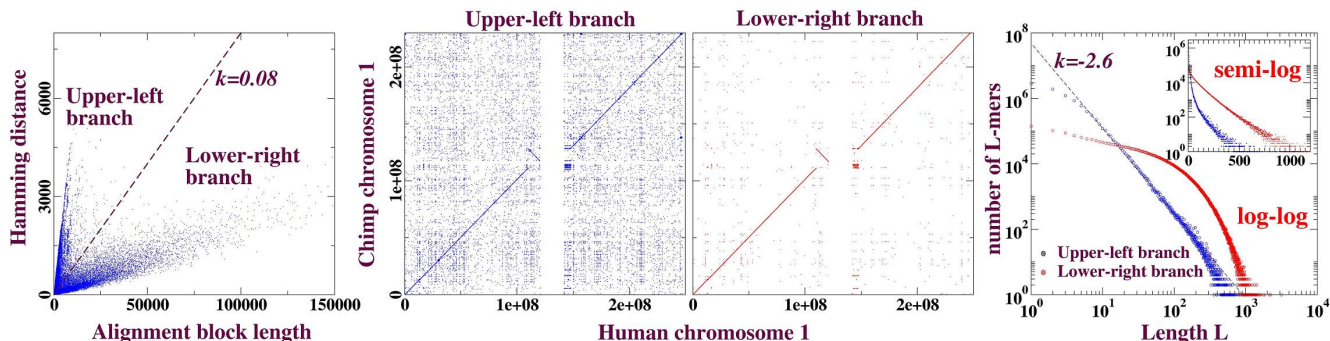


Figure 6: Dot plots and distributions of exact matches depend on the accumulated variation (see text) within the alignment blocks from which they are derived. From raw alignment between human chromosome 1 and chimpanzee chromosome 1, the left panel shows the Hamming distance (number of mismatches and indels) as a function of alignment block length; each dot corresponds to a distinct alignment block. The dashed lines crudely partition alignment blocks into an upper-left branch and a lower-right branch. Dot plots of exact matches for alignment blocks within each branch are shown in the middle panels, and their respective distributions in the right panel, exhibiting decomposition into (approximately) algebraic and exponential components.

3.2.3. “Global” method: Extracting the net alignment from the raw alignment

LASTZ alignment is performed in stages, with “raw” alignment the immediate product. Raw alignment contains all matches between sequences whose alignment scores exceed a predetermined threshold. Aligned fragments often overlap within the raw alignment; one location in the target sequence can match multiple locations in the query sequence, and vice versa. Net alignment scans the target sequence and selects from the aligned fragments in each region the pair with the highest alignment score, discarding all pairs with lower scores, eliminating overlaps and returning a unique optimal chain of aligned fragments [14].

Since the exponential distribution of CMRs in human-chimpanzee alignment comprises primarily of high-similarity sequence pairs, one would expect the net alignment to extract such pairs from the raw alignment. We define a “raw minus net” (RMN) alignment as the residual of the raw alignment once all fragments also in the net alignment have been removed. Thus the net alignment and the RMN alignment represent complementary subsets of the raw alignment.

Kent et al. designed the net alignment to align orthologous sequences [14], so it is not surprising that the LASTZ net alignment between human chromosome 1 and chimpanzee chromosome 1 consists primarily of exponential components.

Figure 7 exhibits dot plots and distributions of exact matches for raw, net and RMN alignments between human chromosome 1 and chimpanzee chromosome 1; and it can be seen there that the net alignment extracts an exponential component from the raw alignment; the RMN alignment distills an (approximately) algebraic component. Figure *S4* exhibits these plots and distributions separately for forward alignment and for reverse alignment.

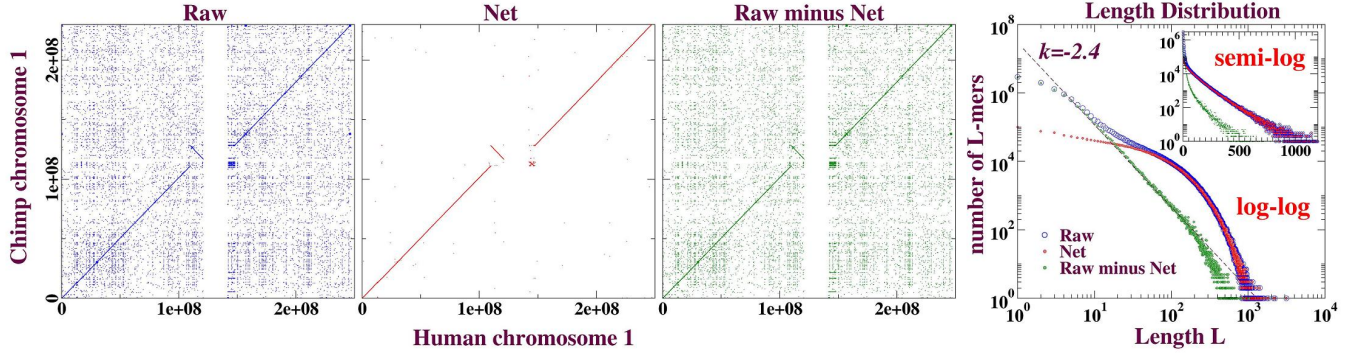


Figure 7: Dot plots and distributions of exact matches for the raw, net and raw minus net alignments between human chromosome 1 and chimpanzee chromosome 1. The net alignment extracts an exponential component from the raw alignment; the RMN alignment distills an (approximately) algebraic component.

3.2.4. “Local” method: Separating non-nested-CMRs from nested-CMRs

We define “nested-CMRs” and “non-nested-CMRs” as two complementary subsets of the CMRs within an alignment: a CMR is said to be “nested” if it is a subsequence of another CMR. In more detail,

Definition 1: If $seq: [i_1, i_2]$ denotes a sequence that starts at location i_1 and ends at location i_2 in a genome (here $i_2 \geq i_1$ are coordinates in the genome, both relative to the plus strand), then for two sequences extracted from a same genome, $seqA: [x_1, x_2]$ and $seqB: [y_1, y_2]$, we say “ $seqA$ is nested in $seqB$ ” if both these conditions are satisfied:

1. $y_2 - y_1 \geq x_2 - x_1$;
2. $y_1 \leq x_1$;
3. $y_2 \geq x_2$;

Definition 2: Given two different CMRs within an alignment, when the query or target sequence of one CMR is nested in the corresponding query or target sequence of the other, we say the first CMR is nested in the second CMR, and the overlap between these two CMRs is called a “nested overlap.”

Definition 3: A CMR that is nested in another CMR is called a *nested-CMR*; otherwise it is a *non-nested-CMR*.

These definitions of nested and non-nested CMR apply to any alignment, including – but not limited to – LASTZ raw [13] and LASTZ net [14] ¹. Below, we apply these definitions to LASTZ raw alignment, and study the distributions exhibited by nested and non-nested CMRs.

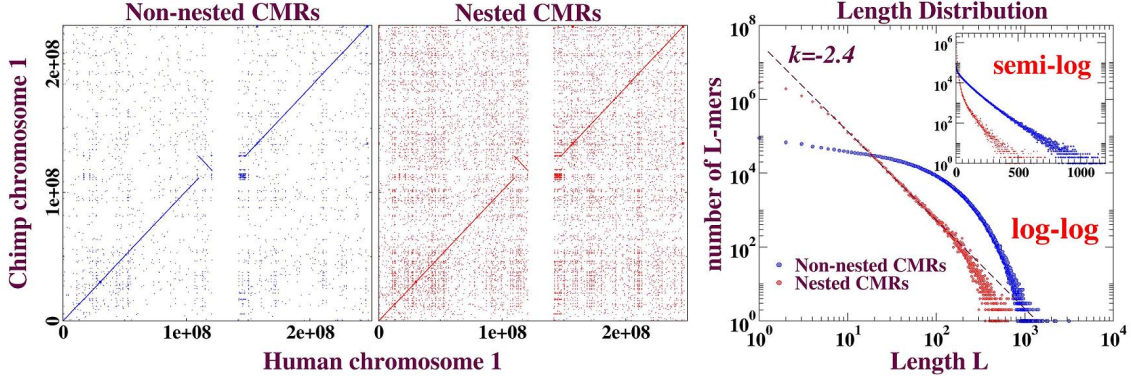


Figure 8: Dot plots and distributions of exact matches for nested-CMRs and non-nested-CMRs in human chromosome 1 – chimpanzee chromosome 1 raw alignment.

Figure 8 exhibits dot plots and distributions for the nested-CMRs and non-nested-CMRs in human chromosome 1 – chimpanzee chromosome 1 raw alignment (for forward and reverse alignments alone see figure S5). The (approximately) algebraic character of nested-CMRs versus the exponential character of non-nested-CMRs is evident. This outcome is plausible if one recalls that orthologs tend to be more similar to one another than are paralogs (see section 4.1), so that subsequences of paralogs are likely to be nested within subsequences of orthologs. This method requires no chaining of alignment blocks, and is further distinguished from netting because it is parameter-free.

3.2.5. Different methods are consistent with one another

Aside from their common reliance on the raw alignment, these four methods (3.2.1 – 3.2.4) are independent of one another; however, the distributions of the corresponding subsets extracted by each of these four methods are

¹However, please note that here our definitions of nested and non-nested CMRs are different from those of nested and non-nested local maxmers by E Tallefer and J Miller in [31].

largely similar. Differences are only apparent in the dot plots. For example, to obtain the exponentially distributed subset, method 3.2.1 extracts the entire diagonal band, discarding all off-diagonal elements. In contrast, the other methods all retain some on-diagonal and some off-diagonal elements.

Figure 9 schematically displays the consistency of these methods, indicating that exponential subsets extracted by different methods consist overwhelmingly of shared CMRs; in particular our “global” and “local” methods share close to 95 ~ 98% of the CMRs (figure 9 left panel).

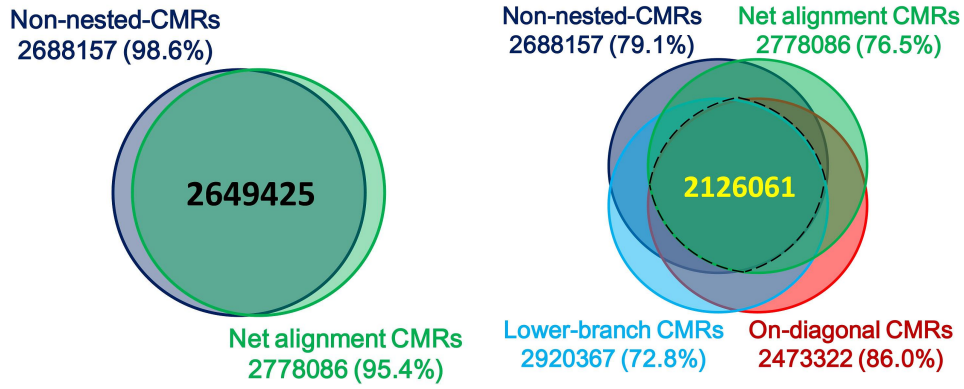


Figure 9: Schematic illustrations of consistency among different methods 3.2.1 – 3.2.4 described in the text. Circles in the figure indicate the exponential subsets extracted from human chromosome 1 – chimpanzee chromosome 1 raw alignment. Numerals in the figure show the number of CMRs in different subsets and percentages in the brackets show the proportions of shared CMRs.

As evident in the right panel of figure 9, the set of CMRs common to all four methods contains at least 70% of the CMRs obtained by each method alone. Although each of these four methods yields some CMRs that are not obtained by any of the other methods, the proportions of such CMRs are small: 1.6% of the net alignment CMRs, 0.4% of the non-nested CMRs, 6.2% of the low-branch CMRs and 7.8% of the on-diagonal CMRs.

3.3. Random uncorrelated point mutation (RUPM) model

To account qualitatively for the ortholog contribution to the exponential distribution, we apply a random uncorrelated point mutation (RUPM) model. As a simple model of neutral evolution, a RUPM model consists of site-independent point mutations (here, single-base substitutions) only, where the rate of these mutations is homogeneous across the genome.

As two identical copies of a common ancestor genome evolve independently under a neutral RUPM model, CMR lengths follow an exponential distribution. For sufficiently short times, long CMRs can be assigned to corresponding positions within the two genomes, and lie on the diagonal; long segmental duplications present in the common ancestor remain well conserved. Matches among these segmental duplications in different locations of the genomes yield a distribution similar to that of the common ancestor: any differences can only be accounted for by random, uncorrelated point mutation.

3.3.1. A synthetic alignment under the RUPM model

We perform a numerical simulation of neutral evolution under the RUPM model. Human chromosome 1 was selected as a common ancestor sequence containing algebraically distributed segmental duplications. Starting from two identical copies of the ancestral genome, random uncorrelated point mutations are introduced independently. We apply 0.5% mutations per base and generate a raw alignment between the descendent genomes. The distributions of exact matches are displayed in figure 10.

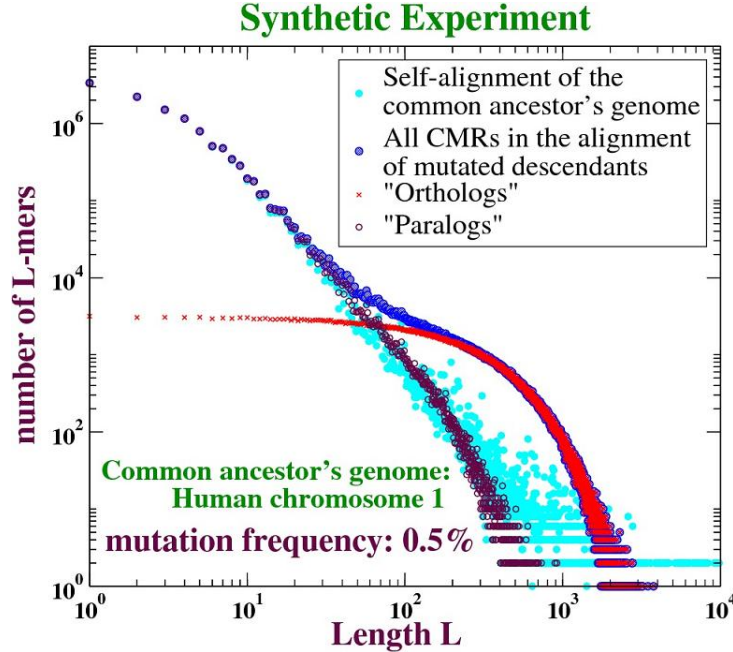


Figure 10: Distributions of exact matches in raw alignment between two “synthetic” descendants of a common ancestor genome. Introducing random uncorrelated point mu-

tations with frequencies 0.5% into two copies of an ancestral genome consisting of human chromosome 1 generates two synthetic descendent genomes. In the figure, solid cyan circles indicate self-alignment of the original (un-mutated) sequence; solid blue circles all the CMRs in the alignment between the mutated sequences; red crosses the “orthologs;” open maroon circles the “paralogs.” “Orthologs” correspond to matches that share common locations between the two descendent genomes’ “paralogs” to matches with a different location in each of the two descendent genomes.

Under the RUPM model, we identify matches between sequences having identical coordinates within the respective mutated sequences as “orthologs” and all other matches as “paralogs.” In figure 10 these orthologs exhibit an exponential distribution, whereas paralogs exhibit an (approximately) algebraic distribution that resembles the algebraic distribution of the self-alignment of the original (un-mutated) sequence, but falls a little short in the tail.

For comparison, a parallel simulation on a random sequence is performed; see **Supplementary Text S1**.

3.3.2. Separating orthologs from paralogs with different methods in the synthetic alignment

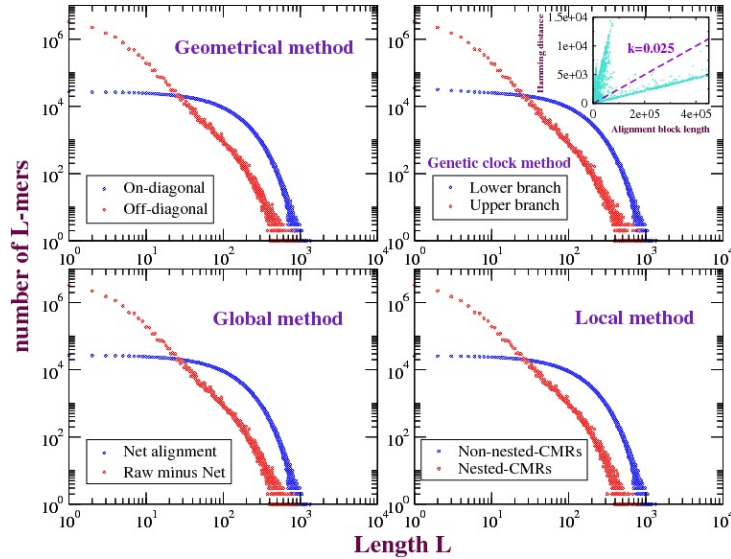


Figure 11: Distributions of the “orthologs” and “paralogs” in our “synthetic” alignment, separated by different methods.

Because evolution is simulated according to the RUPM, the orthologs and paralogs in this synthetic alignment can be identified solely by their lo-

Methods	Subsets	numbers of orthologs	numbers of paralogs	Error (%)
“Geometrical”	On-diagonal	2456358	0	0
	Off-diagonal	0	12169905	0
“Genetic clock” (ratio threshold: 0.025)	Lower branch	2445738	62405	2.49%
	Upper branch	10642	12107501	0.09%
“Global”	Net alignment	2456357	17	0.0007%
	RMN alignment	7	12169893	0.00006%
“Local”	Non-nested-CMRs	2412292	11854	0.489%
	Nested-CMRs	44073	12158052	0.361%

Table 1: Identification of “orthologs” and “paralogs” in the synthetic alignment by methods 3.2.1 – 3.2.4.

cations within the aligned sequences and we can use this synthetic alignment to examine the reliability of the methods 3.2.1 – 3.2.4 above. Figure 11 illustrates the distributions of the “orthologs” and “paralogs” from our synthetic alignment, as separated by each of our four methods; evidently all of them are effective at separating the exponential from the power-law, as can also be seen from Table 1. Relative to the “geometrical” method 3.2.1, which is – *for the RUPM model* – perfect, the other methods also perform well.

3.4. Other orthologous chromosome pairs from human and chimpanzee

The calculations above were performed on human chromosome 1 and chimpanzee chromosome 1. Figure S6 exhibits distributions of exact matches from net and raw minus net alignments of all pairs of orthologous chromosomes from human and chimp. Exponential distributions characterise the net alignments, and algebraic most of the raw minus net. Some chromosome pairs show exponential tails in raw minus net, for example, chromosome 16 and chromosome Y; it happens that these two chromosomes appear to contain more repetitive sequences than other chromosomes (data not shown); however, further understanding awaits future research.

3.5. When species become more distantly related

Heretofore we have addressed only the human-chimpanzee alignment. Whether our conclusions apply equally well to other genome pairs with similar evolutionary distances remains to be seen. Figure 12 shows the distributions of exact matches in alignments between human (*hg19*) chromosome 1

and orthologous chromosomes selected from the Venter, chimpanzee, gorilla, orangutan, macaca and marmoset genomes. We choose for each species the orthologous cognate as the chromosome that shares the most orthologous genes with human chromosome 1 according to Ensembl Biomart (data not shown). For a more distant genome, mouse chromosome 1 is aligned to human chromosome 1; it carries on the order of 1/4 of orthologs between human chromosome 1 and the mouse genome (see e.g. the human-mouse synteny map, <http://cinteny.cchmc.org/doc/wholegenome.php>). As the species pair diverges, distributions of shared sequences gradually cross over from exponential to algebraic. This crossover remains to be accounted for.

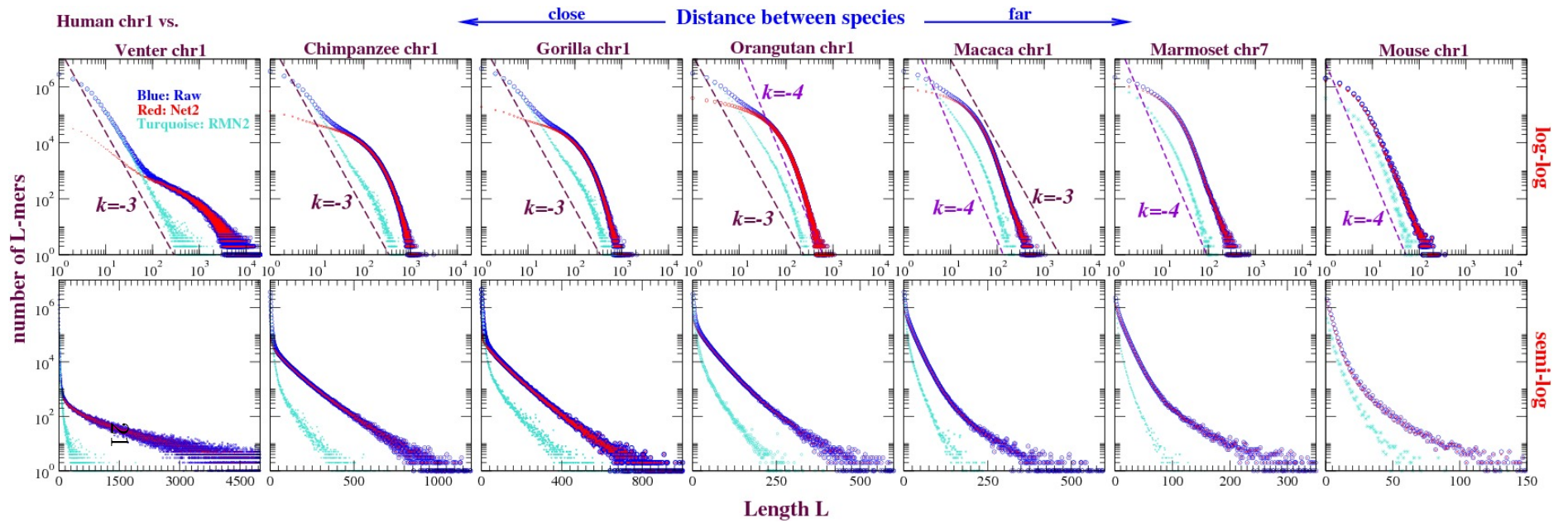


Figure 12: Distributions of exact matches from raw, net and raw minus net alignments of human chromosome 1 versus the corresponding orthologous chromosomes of respectively Venter, chimpanzee, gorilla, orangutan, macaca, marmoset; and in the rightmost panel versus mouse chromosome 1.

4. Discussion

The quantitative study of monoscale substitution/duplication dynamics was revitalised by the work of H.C. Lee and collaborators with their apt characterisation of “nature as the blind plagiariser” [6]. Although these authors did not investigate the steady state duplication length distributions yielded by their models, subsequent research revealed that similar classes of models yield algebraic length distributions that resemble those often exhibited by duplicated sequence in self-alignment and self-intersection of natural genomes [9, 16, 24]. Algebraic distributions of conserved sequence lengths among distantly related genomes had been observed earlier.

This manuscript extends the characterisation of sequence length distributions to a pair of closely related genomes, those of human and chimp, where both conserved sequence lengths and duplicated sequence lengths can be simultaneously computed. In **Results** we demonstrated that the human chromosome 1 – chimpanzee chromosome 1 alignment can be decomposed into two subsets, one with an exponential length distribution, the other an (approximately) algebraic length distribution. Our calculations also suggest that the algebraic length distribution is composed primarily of duplicated sequence including but not limited to paralogous genes, whereas the exponential length distribution is mainly composed of matches between orthologous chromosomal regions.

A neutral substitution model in the absence of selection is expected to yield an exponential length distribution for sequence conserved between two genomes. The phenomenon is quantitatively and conservatively thought of as a Bernoulli process; the exponential arises from the length distribution of head runs when flipping a biased coin [3], and the exponential underlies most null models of sequence similarity in comparative genomics. It is not understood the extent to which an exponential is expected in (say) human/chimpanzee alignment, or whether – since we are not chimpanzees – an exponential is unexpected because of selection.

4.1. Orthology and paralogy

Chromosomal regions, within or across species, that have common ancestry are said to be *homologs* [4]. Homologs can be further identified as *orthologs* if they diverged via evolutionary speciation, or *paralogs* if they diverged via sequence duplication [2, 7]. Orthology and paralogy can in principle be defined for all sequences within a genome, but in practice most on-line

databases consist only of protein-coding genes. Because of gene duplication and genome rearrangement, the ancestry of a given gene may be difficult to ascertain with high confidence, and ortholog/paralog classification can be ambiguous. Phylogenetic analysis of the gene lineage is customarily believed to enable the strongest discrimination between orthology and paralogy.

A standard approach to orthology and paralogy is to argue that within a given genome pair, orthologs tend to be those homologs that diverged least [2]. Duplication subsequent to speciation generates “mother” and “daughter” copies (known as “in-paralogs”) that exhibit congruent divergence from their cognate orthologs. This sequence of events yields so-called “co-orthology” among in-paralogs [19]. Co-orthology can be further refined to “primary orthology” and “secondary orthology” [12]. Our preliminary calculations suggest that in the human-chimpanzee alignment, *primary* orthologs dominate the exponential length distribution, but *secondary orthologs* merge with paralogs into the power-law length distribution.

4.2. Approximate matching

In the plots above we study continuous (uninterrupted) matching runs of bases (CMRs), where continuous matching runs are by definition terminated at mismatches or indels; these are exact matches; however, CMRs may also be defined according to approximate matching criteria. The following criteria are listed in order of decreasing stringency:

- I : Exact matches: Each of the four nucleotides (A,T,G,C) matches itself only; a mismatch or indel terminates a run of matches;
- II : A=G, C=T: In addition to the exact matches, A and G, C and T are also matched pairs; an indel or any mismatch involving other than an A/G or T/C pair terminates the run;
- III : Indel-terminated matches: aligned but gap/insert-free sequence is taken as matching; only an indel terminates the run;
- IV : Alignment blocks: High similarity local alignments returned by LASTZ that are separated from one another by un-alignable sequence. They span exact matches, mismatches and indels.

CMR distributions obtained with criteria I through IV display sufficient qualitative similarity to one another that only exact match distributions are

displayed in this manuscript. An example for human-chimpanzee alignment can be found in the supplement (see figure *S9* in **Supplementary Text S2**); for other genome pairs corresponding plots may be found in [9, 17, 26, 30, 31].

It was observed for distant *inter*-genome comparisons in [30] and [26] that criterion II matches – in contrast to all other inexact base substitution matching conditions – displace the algebraic distribution of exact matches to numbers and lengths greater by an order-of-magnitude, with minimal impact on the shape of the curve. Were these $C \Rightarrow T / G \Rightarrow A$ substitutions neutral, an exponential would have been anticipated. Yet, a qualitatively similar phenomenon (criterion II shifts algebraic criterion I curves to larger numbers and greater lengths, with minimal impact on shape) is observed for duplications *within* a genome [9, 16, 31].

4.3. A conjecture on the crossover of orthologous sequence from exponential to algebraic

The qualitative parallels between distributions of exact and inexact matches in duplicated sequence versus conserved sequence – discussed in the previous section – suggest to us that the mechanisms behind them share common features. Subsequent to our original computations [26, 30], the portfolio of fully sequenced genomes has expanded vastly, and a variety of genome pairs exhibiting exact match length distributions with power laws close to -3 have emerged (jm, unpublished). This leads one of us (jm) to conjecture, supported by preliminary numerical calculations, a class of models that can account qualitatively for these observations.

The proposed class of models builds on the notion of sequence dynamics as fragmentation of a steady source [17, 18, 20, 24, 33]. The biological realisation of a mean steady source of newly duplicated sequence is readily plausible; its counterpart for sequence conservation may be more speculative. For sequence conservation, it is suggested that the counterpart of duplication is the steady generation of novel constrained sequences on which the constraints are newly relaxed. A few years ago, this notion seemed implausible, as the consensus among most biologists was that new functionalities arise through sequence duplication; however, recently evidence has emerged for alternative routes [5, 29]. How much sequence arises through these alternative routes – and what constitutes them – is still unclear; for our purposes, it is not necessary to be too specific about details of any mechanism. Rather, we regard adaptation on the sequence level as a process of steady production (over evolutionary timescales) of novel sequence that serves novel functionalities, coupled with

relaxation of or loss of constraint on sequences whose functionalities have become obsolete.

The latter yields a steady source of newly unconstrained sequence in the common ancestors that is reflected in descendants by randomly fragmented subsequences, as indicated in figure 13. In figure 13 (a), the opaque coloured blocks represent newly duplicated sequence within a single lineage. The fading colour indicates the loss of homology between a duplicate sequence and its source as random local mutations fragment the matches. The time elapsed between the given duplication event and the present, governs the extent of fragmentation of the given duplicate. In figure 13 (b), the opaque coloured blocks represent sequence – not necessarily duplicated – on which selection has been newly lost. The faded colour indicates the loss of homology over time, as the newly unconstrained sequence accumulates random local mutations that fragment the matches. The evolutionary distance between a pair of genomes at the leaves of the tree – reflecting the time elapsed between loss of constraint on the given sequence and the present – governs the extent of fragmentation of the given sequence.

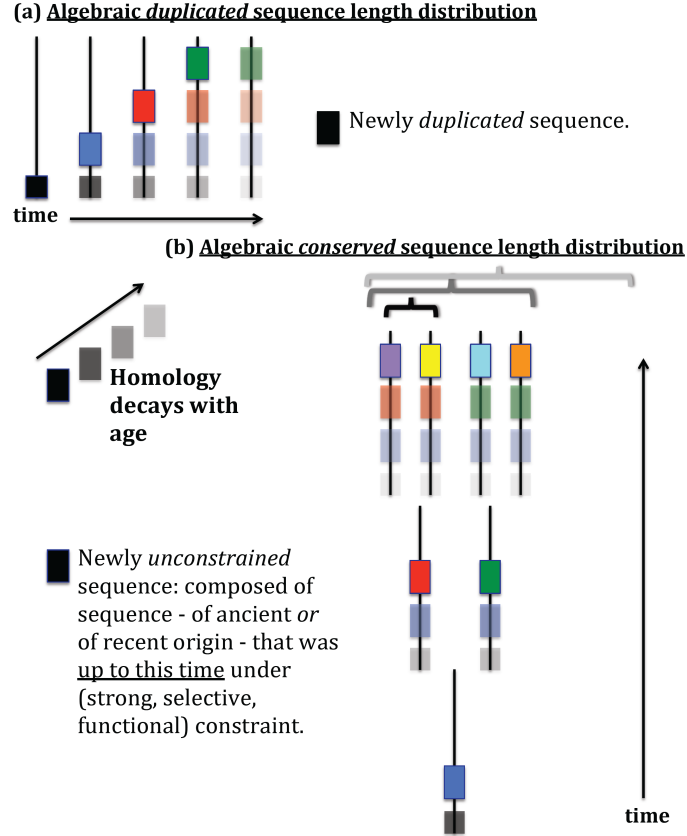


Figure 13: Schematic illustration of how a steady source of homologous sequence, subject to local mutation such as random base substitution, can lead to stationary algebraic distributions of homologous sequence length. (a) Solid coloured blocks represent newly duplicated sequence within a single lineage. The fading colour indicates the loss of homology between a duplicate sequence and its source as random local mutations fragment the matches. The time elapsed between the given duplication event and the present, governs the extent of fragmentation of the given duplicate. (b) Solid coloured blocks represent sequence – not necessarily duplicated – on which constraint has been newly lost. The faded colour indicates the loss of homology over time, as the newly unconstrained sequence accumulates random local mutations that fragment the matches. The time elapsed between the loss of constraint on the given sequence and the present, governs the extent of fragmentation of the given sequence. The braces indicate how this time is reflected in evolutionary distance: nearby leaves (black brace) are dominated by recent mutations of unconstrained sequence and yield an exponential distribution; intermediate distances (dark grey brace) are dominated by an algebraic distribution arising from successive losses of constraint; at still greater distances (light grey brace) the distribution exhibits increasingly steep tails as the overall amplitude attenuates into noise.

When comparing a *pair* of present-day descendants of a common ancestor, fragmentation could be misinterpreted as representing the *average* constraint on the sequence over all time; sequences that lost their constraints earlier appear subject to less constraint overall (are more fragmented) than sequences that lost their constraints more recently. Presumably, only suitable outgroup genomes can resolve this potential ambiguity.

Observe that, in accord with figure 13 (b), recently diverged sequences (nearby branches) are expected to share exponentially distributed exact match lengths (because all the mutations breaking the matches occurred *subsequent* to divergence); an intermediate regime to share algebraically distributed match lengths (the integral of fragment lengths arising from mutations that occurred *before* divergence), in principle with power -3 ; as the branches separate further the amplitude of the distribution diminishes until matches are too sparse to infer its form.

In summary, the parallel between the algebraic distributions of duplicated and conserved sequence is that they both represent a signature of perpetual sequence *turnover*; for conserved sequence, the *turnover of functional sequence* in a continual process of expropriation, exploitation, and extinction. The latter conception is hardly novel, but the prospect of a quantitative measure of it (the exponent, presumably) could be illuminating.

5. Conclusion

Exponential length distributions between similar species and algebraic (power-law) length distributions between more distantly related species and within the alignment of a genome to itself have been previously observed. We have studied here the distribution of lengths of identical (and nearly identical) sequence shared between closely related organisms. A key contribution of our study is that the exponential distribution between closely related genomes turns out to be composed of two types of sequences: (1) orthologous sequences, which have an exponential distribution; (2) paralogous sequences, which have an algebraic (power-law) distribution.

Comparing human and chimpanzee, we explicitly distinguish orthologous from non-orthologous regions in a number of different ways, including known chromosome orthology; annotated orthologous regions in chromosomes; diagonal versus non-diagonal sectors of a dot plot; alignment similarity between human and chimpanzee; optimal chains of fragments aligned between orthologous chromosomes. For all such characterisations, we demonstrate expo-

nentially distributed length segments for orthologous regions, and algebraic (power-law) distributed length segments for non-orthologous regions. Finally, we provide an *in silico* demonstration of how such length distributions could have arisen through neutral evolution.

Recent models of neutral evolution proposed to explain algebraic distributions of duplicated sequence lengths often observed in natural genomes lead one to ask whether they can shed light on the evolution of duplications over evolutionary time scales [17, 24]. Addressing this question suggests the investigation of duplicated sequences common to at least two different species. At the same time, observations from almost ten years ago of algebraic distributions of sequences conserved among multiple divergent species remain unaccounted for [30].

In this paper, we take some first steps in studying the evolution of the distribution of duplicated sequence lengths from self-alignment to alignment of two nearby species, human and chimpanzee. We describe a parameter-free method of extracting paralogs from LASTZ raw alignment of human and chimpanzee, based on nested and non-nested matches, that seems to reconstitute an approximately algebraic distribution of shared duplicate sequence lengths traceable to the self-alignment. Finally, we exhibit the evolution of orthologous sequence length distributions over a range of increasingly divergent species that spans the exponential and the algebraic, for which a mechanism is conjectured.

As observed in [30] (twenty years after the rigorous mathematics of [3]) pure exponentials may not be so easy to come by in natural genome sequences. Once that has been recognized, the relevant question shifts to “under what circumstances do exponentials actually occur, and why or why not?” And if not, what takes their place and what does it tell us about sequence evolution? We hope that the work presented here will eventually lead to further insights into these questions.

Acknowledgement

The authors gratefully acknowledge generous support from the Okinawa Institute of Science and Technology Graduate University to jm.

References

References

- [1] A Albrecht-Buehler. Inversions and inverted transpositions as the basis for an almost universal “format” of genome sequences. Genomics, 90:297–305, 2007.
- [2] AM Altenhoff and C Dessimoz. Inferring orthology and paralogy. In M Anisimova, editor, Evolutionary Genomics: Statistical and Computational Methods, chapter 9. Springer Science+Business Media, 2012.
- [3] R Arratia and MS Waterman. Critical phenomena in sequence matching. Annals of Probability, 13(4):1236–49, 1985.
- [4] TA Brown. Molecular phylogenetics. In Genomes, 2nd Edition, chapter 16. Oxford: Wiley-Liss, 2002.
- [5] JA Capra, KS Pollard, and M Singh. Novel genes exhibit distinct patterns of function acquisition and network integration. Genome Biology, 11(R127), 2010.
- [6] HD Chen, WL Fan, SG Kong, and HC Lee. Universal global imprints of genome growth and evolution: equivalent length and cumulative mutation density. PLoS ONE, 5(4)(e9844), 2010.
- [7] W Fitch. Distinguishing homologous from analogous proteins. Syst Zool, 19(2):99–113, 1970.
- [8] KJ Fryxell and E Zuckerkandl. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol. Biol. Evol., 17(9):1371–1383, 2000.
- [9] K Gao and J Miller. Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments. PLoS One, 6(7)(e18464), 2011.
- [10] AJ Gibbs and GA McIntyre. The diagram, a method for comparing sequences. its use with amino acid and nucleotide sequences. Eur. J. Biochem, 16:1–11, 1970.

- [11] D Grauer and WH Li. Fundamentals of Molecular Evolution. Sinauer, 2000.
- [12] M V Han, J P Demuth, C L McGrath, C Casola, and M W Hahn. Adaptive evolution of young gene duplicates in mammals. Genome Research, 19:859–867, 2009.
- [13] RS Harris. Improved pairwise alignment of genomic DNA. PhD thesis, The Pennsylvania State University, 2007.
- [14] W J Kent, R Baertsch, A Hinrichs, W Miller, and D Haussler. Evolutions cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA, 100(20):11484–11489, 2003.
- [15] EV Koonin and YI. Wolf. The common ancestry of life. Biology Direct, 5(64), 2010.
- [16] MV Koroteev and J Miller. Scale-free duplication dynamics: a model for ultraduplication. Phys. Rev. E., 84(061919), 2011.
- [17] MV Koroteev and J Miller. Fragmentation dynamics of dna sequence duplications. arXiv: 1304.1409v3 [math-ph], 2013.
- [18] PL Krapivsky, S Redner, and E Ben-Naim. A Kinetic View of Statistical Physics. Cambridge University Press, 2010.
- [19] D M Kristensen, Y I Wolf, A R Mushegian, and E V Koonin. Computational methods for gene orthology inference. Briefings in Bioinformatics, 12(5):379–391, 2011.
- [20] W Kuhn. Uber die kinetik des abbaues hochmolekularer ketten. Ber. Dtsch. Chem. Ges., 63(1530), 1930.
- [21] S Levy, G Sutton, PC Ng, L Feuk, AL Halpern, and et al. The diploid genome sequence of an individual human, doi:10.1371/journal.pbio.0050254. PLoS Biology, 5(10), e254, 2007.
- [22] G Lunter, CP Ponting, and J Hein. Genome-wide identification of human functional dna using a neutral indel model. PLoS Comput. Biol., 2(1)(e5), 2006.
- [23] M Lynch. The Origins of Genome Architecture. Sinauer, 2007.

- [24] F Massip and Arndt P F. Neutral evolution of duplicated dna: An evolutionary stick-breaking process causes scale-invariant behavior. Physical Review Letters, 110(148101), 2013.
- [25] E H McConkey. Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. Cytogenetic and Genome Research, 105:157–158, 2004.
- [26] J Miller. Colossal and super-colossal ultraconservation. IEICE Technical Report, Neurocomputing, 109(53), 2009.
- [27] DL Nelson and MM Cox. Lehninger Principles of Biochemistry (6th Edition). W.H. Freeman, 2012.
- [28] S Ohno. Evolution by Gene Duplication. Springer-verlag, New York. Heidelberg. Berlin, 1970.
- [29] CP Ponting, C Nellaker, and S Meader. Rapid turnover of functional sequence in human and other genomes. Annu. Rev. Genom. Human Genet., 12:275–299, 2011.
- [30] W Salerno, P Havlak, and J Miller. Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments. Proc Natl Acad Sci USA, 103(35):13121–13125, 2006.
- [31] E Taillefer and J Miller. Exhaustive computation of exact duplications via super and non-nested local maximal repeats. J Bioinform Comput Biol., 12(1)(1350018), 2014.
- [32] G Varani and WH McClain. The $g \times u$ wobble base pair: A fundamental building block of rna structure crucial to rna function in diverse biological systems. EMBO Rep., 1(1):1823, 2000.
- [33] RM Ziff and ED McGrady. The kinetics of cluster fragmentation and depolymerisation. J. Phys. A, 18(3027), 1985.
- [34] E Zuckerkandl and LB Pauling. Molecular disease, evolution, and genic heterogeneity. Horizons in Biochemistry. Academic Press, New York., 1962.

Supplementary figures

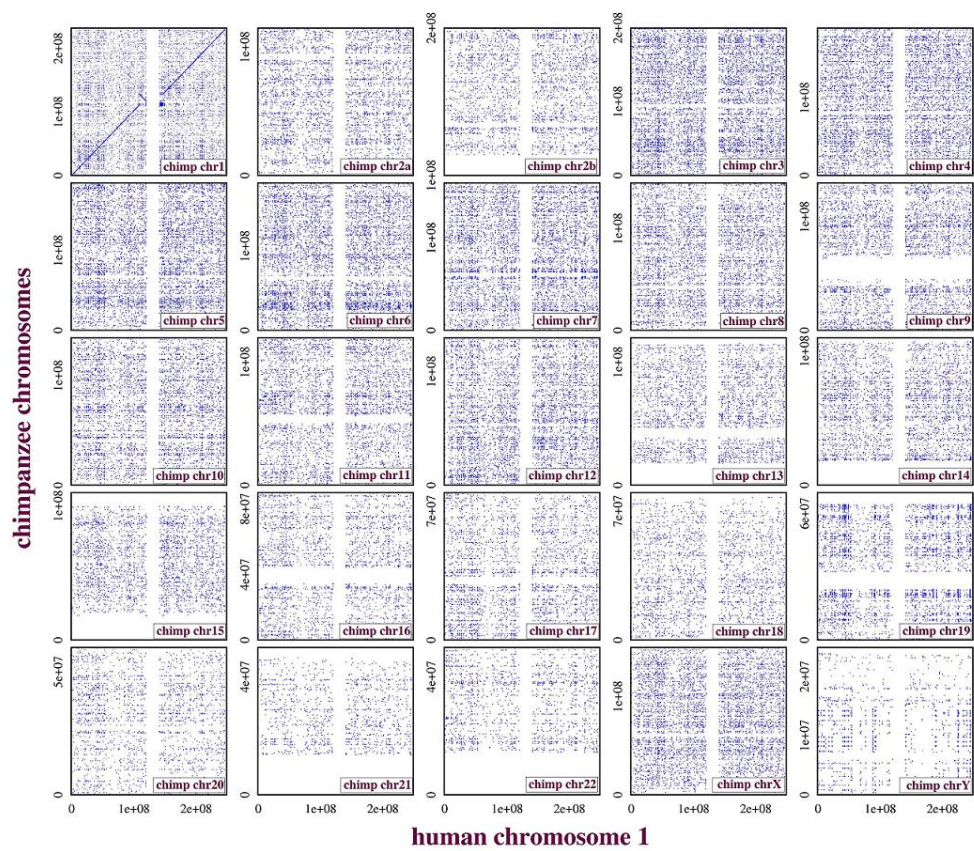


Figure S1: Dot plots for (soft repeat-masked LASTZ) raw alignments between human chromosome 1 and each chimpanzee chromosome.

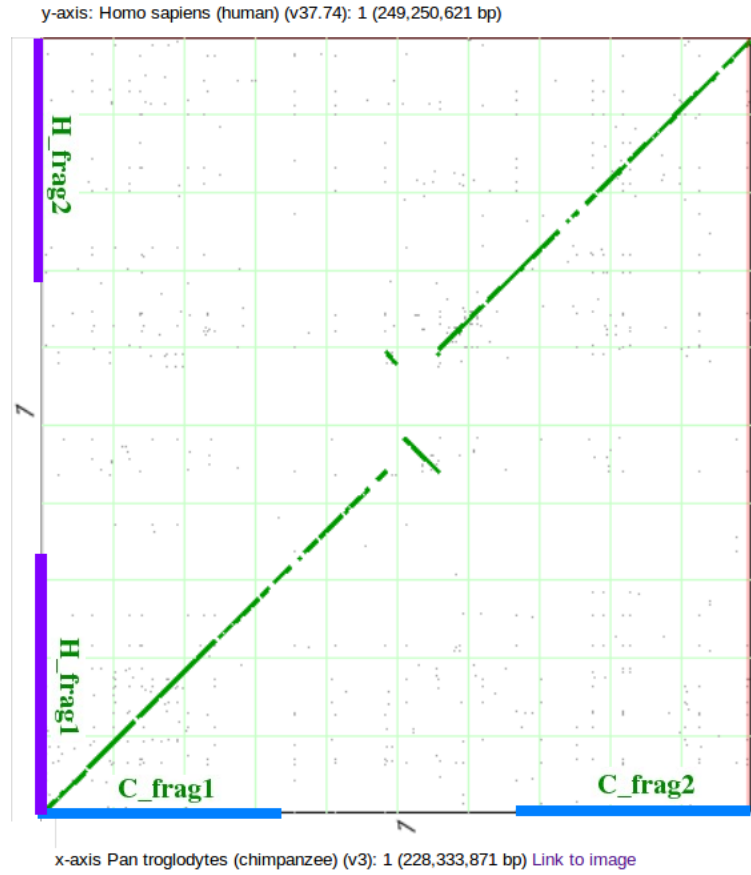


Figure S2: Syntenic dot plot for the CDS (protein-coding nucleotide sequences) between human chromosome 1 and chimpanzee chromosome 1, created by the SynMap tool in CoGe (<http://genomevolution.org/CoGe/index.pl>). The horizontal blue and vertical violet bars indicate the locations of fragments *H_frag1*, *H_frag2*, *C_frag1* and *C_frag2* defined in figure 3.

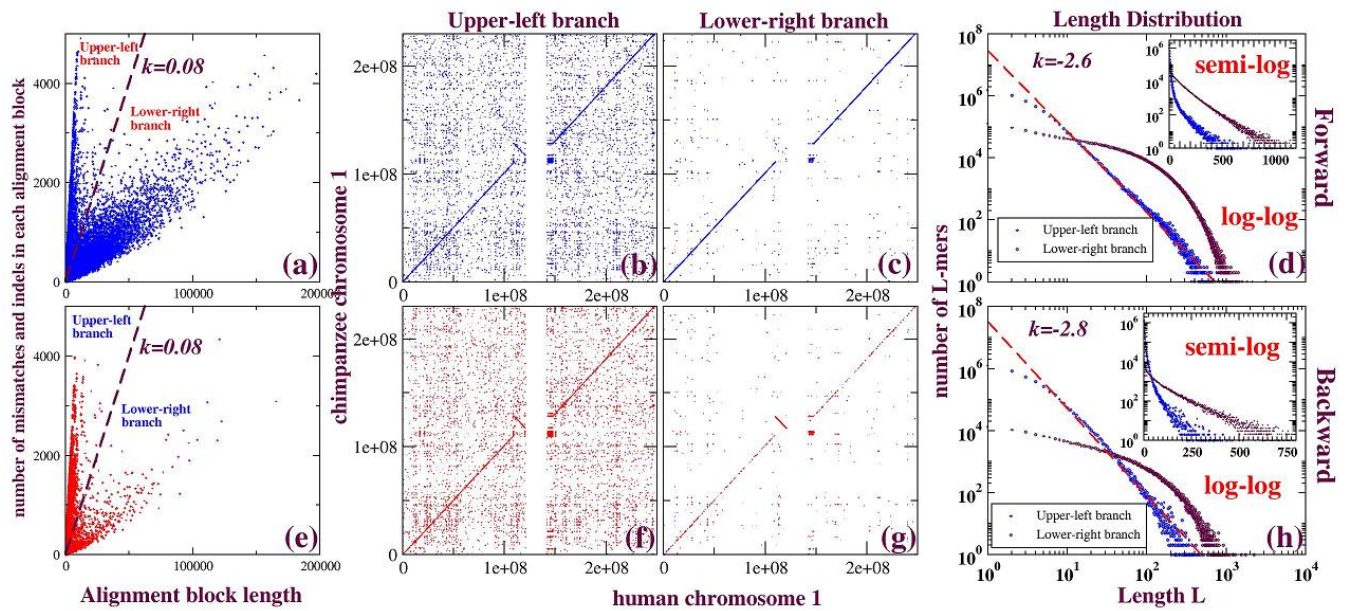


Figure S3: Same as figure 6 in the main text, but displaying separately the forward and reverse alignments.

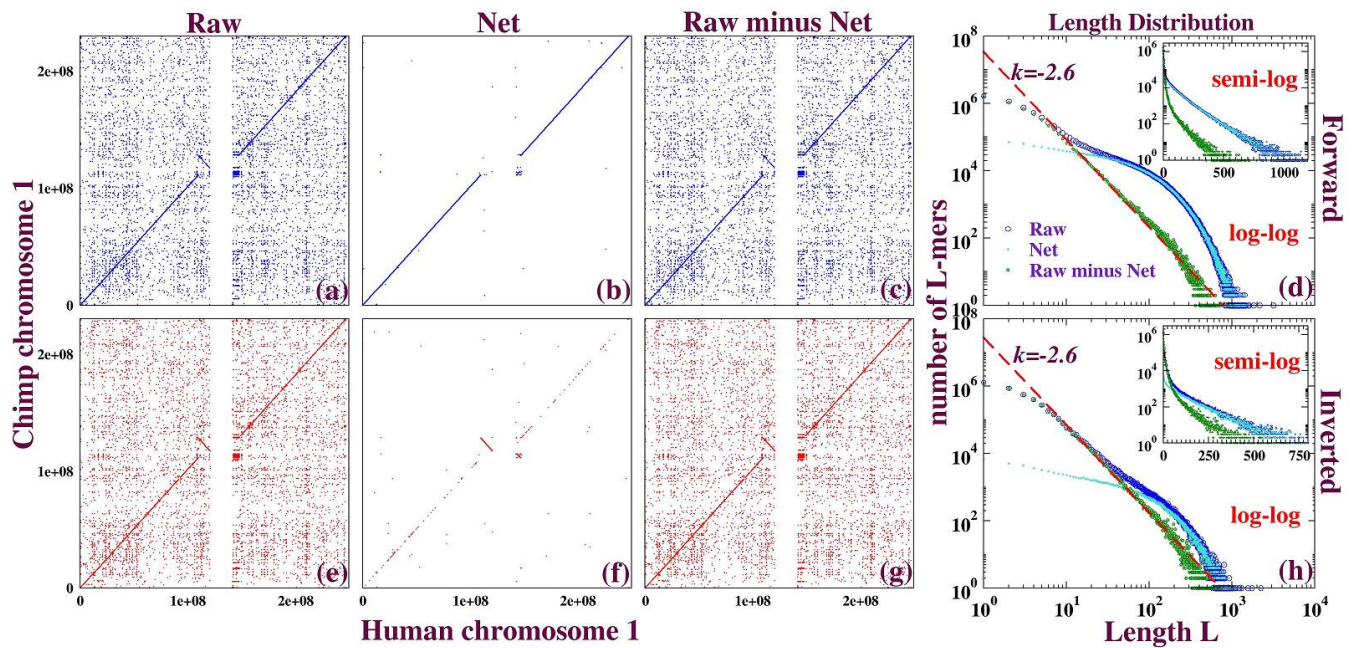


Figure *S4*: Same as figure 7 in the main text, but displaying separately the forward and reverse alignments.

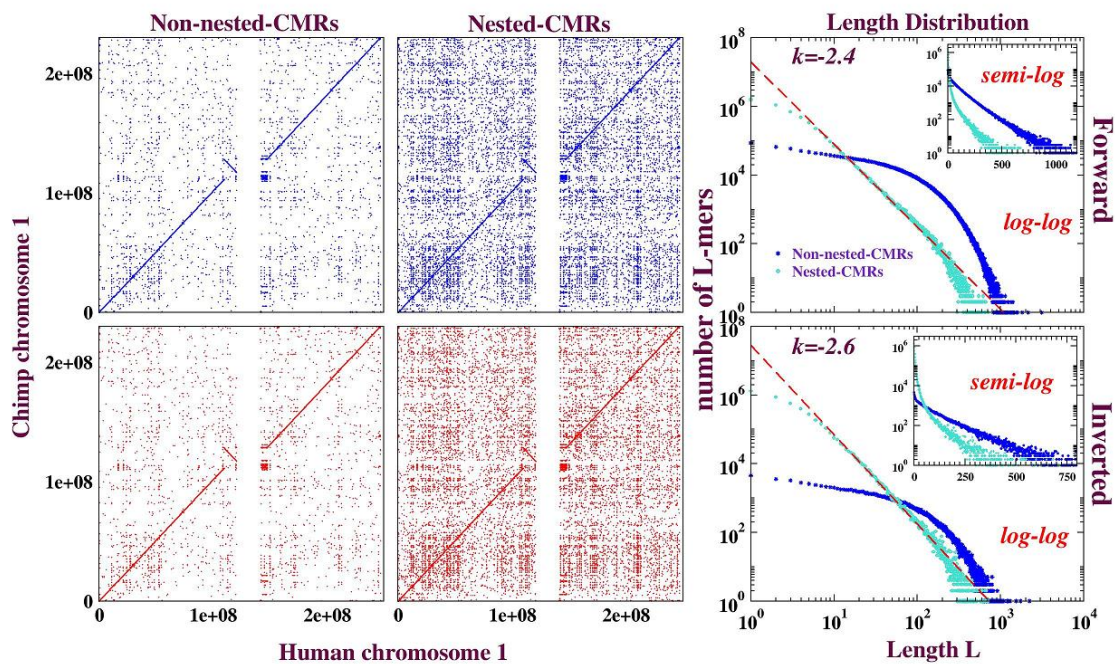


Figure *S5*: Same as figure 8 in the main text, but displaying separately the forward and reverse alignments.

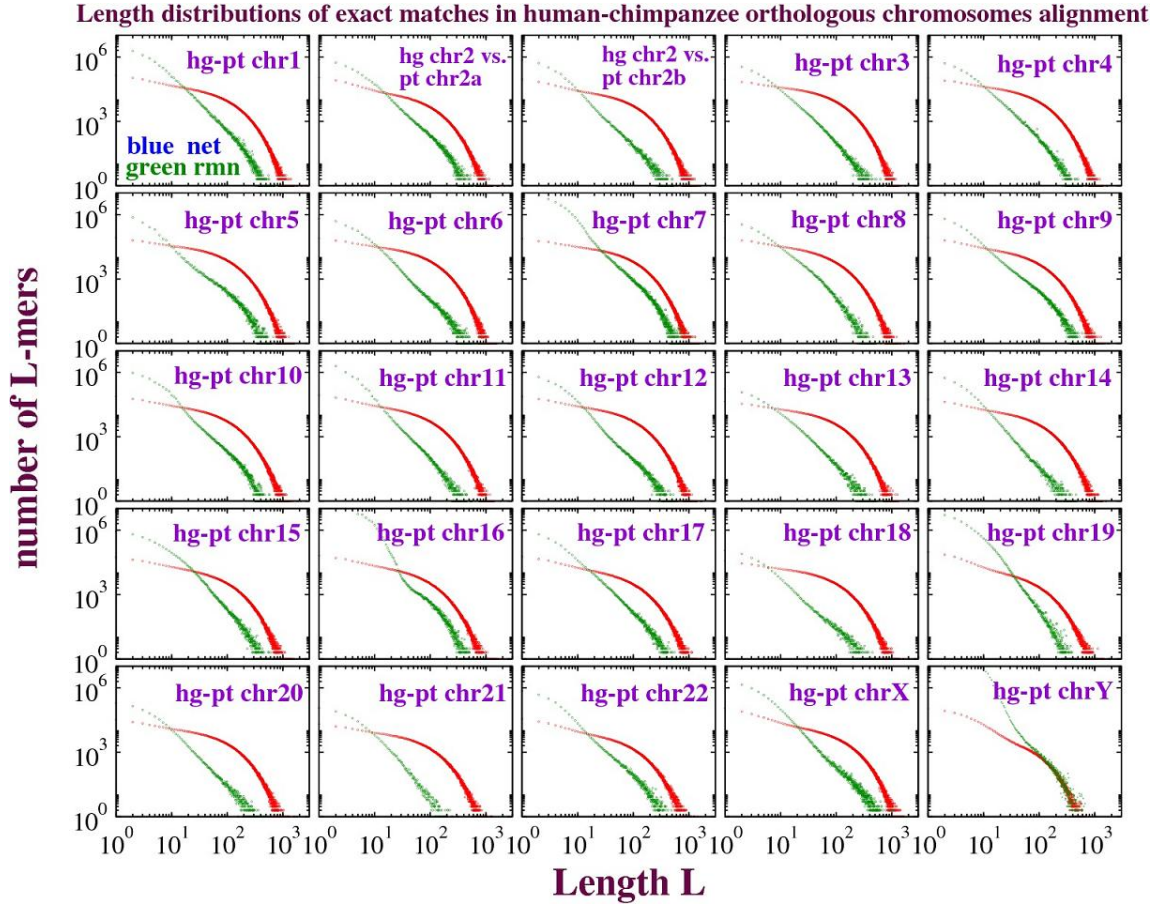


Figure S6: (Length) distribution of exact matches in the net and RMN (raw minus net) alignments of chromosomes orthologous between human and chimpanzee.

Supplementary Text

Supplementary Text S1. Control simulation for the synthetic alignment in section 3.3.1

At the insistence of one of the referees, to confirm our interpretation of the simulation in section 3.3.1 we applied the numerical procedure described there to a random sequence of the same length as human chromosome 1 and with lower-case letters at the same positions as they appear in soft-masked human chromosome 1. Our simulation preserved the case of each letter, because unless this soft-masking was maintained, LASTZ was unable to complete an alignment of the descendent genomes. We applied 0.5% substitutions

per base independently to each of two copies of the random sequence and generated a LASTZ raw alignment between the mutated descendent genomes. “Orthologs” and “other matches” were identified via each of the methods 3.2.1 – 3.2.4. The distributions of exact matches are displayed in figure *S7*.

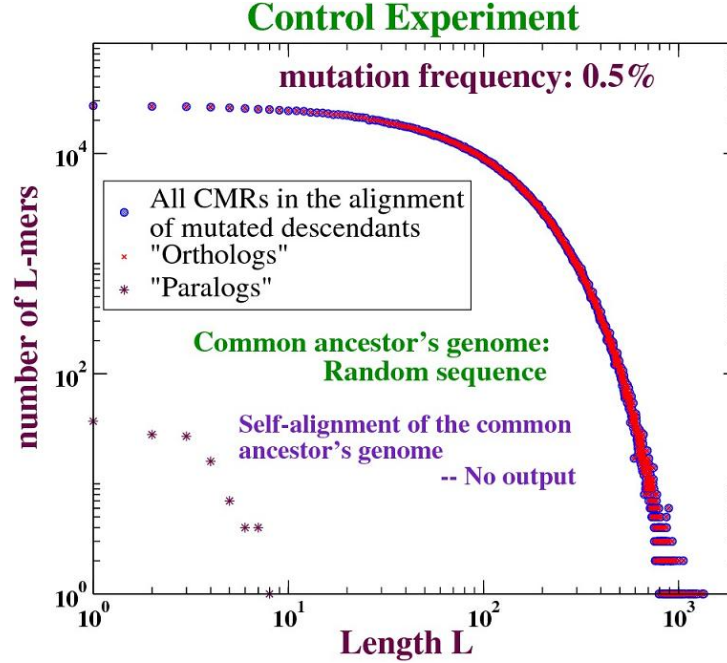


Figure *S7*: A control simulation to figure 10. Random uncorrelated point mutations at a rate of 0.5% per base are applied to produce two independent realisations of the RUPM on a randomly generated “ancestral genome” of the same length as human chromosome 1 and that are soft-masked at the same locations as in human chromosome 1. Since the ancestral genome consists solely of independent uncorrelated random sequence, a power-law distribution of paralogs does not appear in this control simulation.

Figure *S8* and table *S1* for the control simulation correspond respectively to figure 11 and table 1 of the text. “Orthologs” in the synthetic and control simulations share qualitatively similar features, but because the random ancestral genome contains no segmental duplications, paralogs are absent in the control simulation (see inset in upper-right panel of figure *S8*, where the upper-left branch is missing). In figure *S7* any non-orthologous matches appear only by chance.

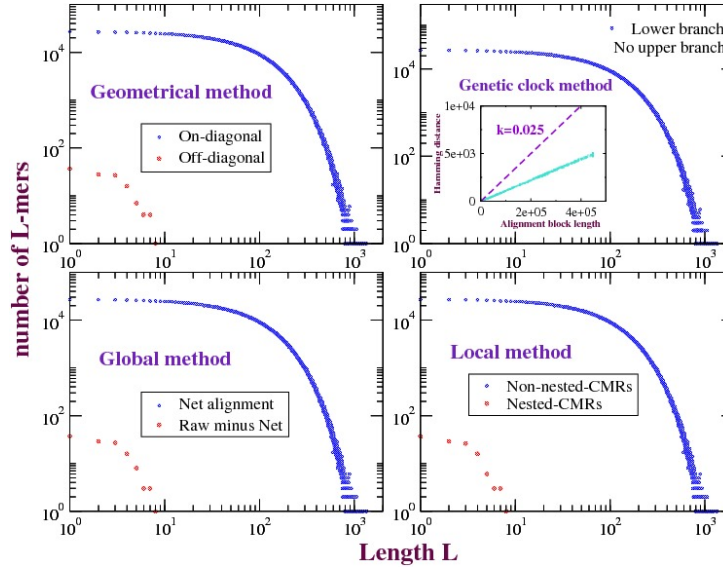


Figure S8: Distributions of “orthologs” and “non-orthologous matches” in the control simulation as identified by each of the methods 3.2.1 – 3.2.4.

Methods	Subsets	numbers of orthologs	numbers of non-orthologs	Error (%)
“Geometrical”	On-diagonal	2454994	0	0
	Off-diagonal	0	124	0
“Genetic clock” (ratio threshold: 0.025)	Lower branch	2455118	0	0.005%
	Upper branch	0	0	0
“Global”	Net alignment	2454994	2	0.00008%
	RMN alignment	2	122	1.613%
“Local”	Non-nested-CMRs	2454993	4	0.0002%
	Nested-CMRs	1	120	0.826%

Table S1: Identifications of orthologs and paralogs in the control simulation by methods 3.2.1 – 3.2.4.

Supplementary Text S2. *Exponential distributions of CMRs counted by different matching criteria*

Matching criteria I through IV described in section 4.2 successively relax the matching condition. CMRs counted according to a stricter criterion are

contained within those counted according to a more relaxed criterion; therefore, locally and within an alignment block, CMRs counted according to different criteria exhibit a nested or hierarchical structure. Different matching criteria yield qualitatively similar length distributions, which suggests to us that the latter reflect some intrinsic features of the genomes rather than artefacts of matching criteria.

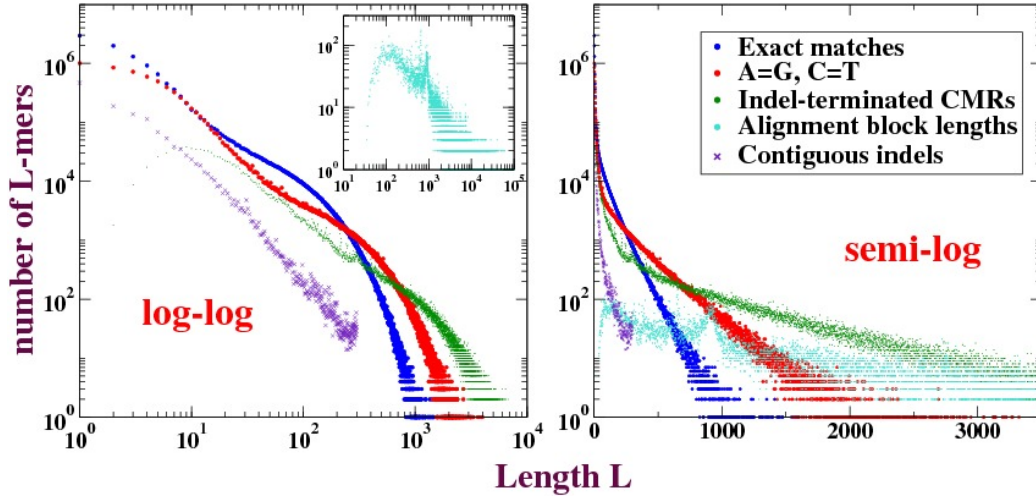


Figure S9: Distributions of the contiguously matched runs counted by different matching criteria in the (soft repeat-masked LASTZ) raw alignment between human chromosome 1 and chimpanzee chromosome 1.

Figure S9 shows the distributions of CMRs counted by different matching criteria from the human chromosome 1 – chimpanzee chromosome 1 LASTZ raw alignment. Evidently, exact matches, $A=G/C=T$ runs and indel-terminated runs all show exponential tails. As described in [9], contiguous indels (successive insertions or deletions) yield algebraic length distributions.

The biological significance of matching criterion II, transition ($G \Leftrightarrow A$, $C \Leftrightarrow T$), has been recognised since the discovery of the genetic code in the mid 20th century [27]. In the genetic code, it is evident that the 3rd base “wobble” displays enhanced tolerance for transitions (they tend not to alter the amino acid encoded by a codon) over other kinds of substitutions. Furthermore, in RNA (DNA) secondary structure, the $G:U$ ($G:T$) base-pair hydrogen bond plays a central role in stabilising duplex structures formed by all classes of RNAs [32]. Thus the tolerance in these contexts for $G \Leftrightarrow A$ and $C \Leftrightarrow T$ substitutions is reflected by functional selection.

On the other hand, the $C \Rightarrow T$ mutation rate (and via subsequent mismatch repair, the $G \Rightarrow A$ rate) is enhanced by an order of magnitude over other substitutions by the (selectively neutral) chemical process of deamination [11]. The effective pressure for $C \Rightarrow T/G \Rightarrow A$ substitution is so strong that it is believed that certain specific mechanisms, such as $A \Rightarrow G$ biased gene conversion, may have evolved to compensate for it [8].

The biological relevance of criterion III (indel terminated) is also widely recognised; consider, for example, the impact of an out-of-frame shift within protein-coding sequence. miRNAs are generally claimed to be insensitive to insertions or deletions within their interiors, but quite sensitive to insertions or deletions in their “seed” regions; the latter observations have been applied quantitatively in [22].